# Machine Learning & Data Mining

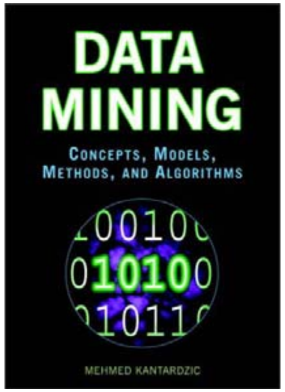
Berlin Chen 2004

References:

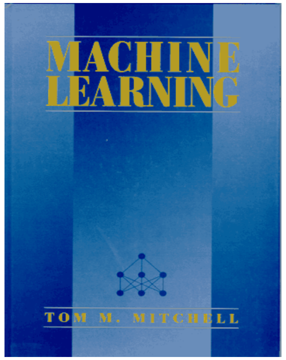
1. Data Mining: Concepts, Models, Methods and Algorithms, Chapter 1
2. Machine Learning , Chapter 1
3. The Elements of Statistical Learning; Data Mining, Inference, and Prediction , Chapter 1
4. Data Mining: Concepts and Techniques , Chapter 1

# Textbooks

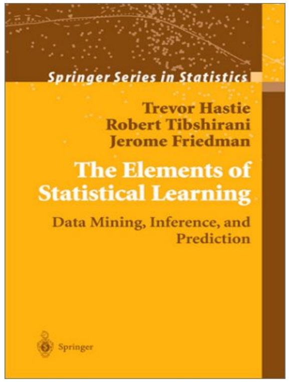
1. Mehmed M. Kantard, Data Mining: Concepts, Models, Methods and Algorithms, Wiley-IEEE Press, 2002




2. Tom M. Mitchell, Machine Learning, McGraw-Hill, 1997




3. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning; Data Mining, Inference, and Prediction, Springer-Verlag, 2001



# References


1. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001
2. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley Longman, 1999
3. I. H. Witten and E. Frank, Data Mining, Morgan Kaufmann, 2000.
4. Stuart Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, Prentice-Hall, 2003
5. Nils J. Nilsson, Artificial Intelligence: A New Synthesis, Morgan Kaufmann, 1998

# Goal

- Know the basic concepts and fundamentals of machine learning and data mining
- Theoretically understand a variety of algorithms that can be used in the fields such as data mining, information retrieval, pattern recognition, …

# Machine Learning

- Address the question of how to build computer programs that improve their performance at some task through experience
  - Learning is a process $\rightarrow$ algorithm/program
- Can be viewed as searching a very large space of possible hypotheses to determine one that best fits the observed data and any prior knowledge held by the learner, and also can correctly generalize to unseen examples
  - Search strategies
  - Underling structures of the hypothesis space

Different learning methods searching different hypothesis spaces

# Why Machine Learning

- Recent progress in algorithms and theory
- Growing flood of online data
- Computational power is available
- Budding industry

# Niches for Machine Learning

- Data mining
  - E.g., using historical data to improve decisions
    - medical records $\rightarrow$ medical knowledge
- Software applications
  - autonomous driving
  - speech recognition
- Self customizing programs
  - Newsreader that learns user interests

# Example: Credit Risk Analysis

## Data:

| *Customer103:* (time=t0) | *Customer103:* (time=t1) | ... | *Customer103:* (time=tn) |
|---|---|---|---|
| Years of credit: 9 | Years of credit: 9 | | Years of credit: 9 |
| Loan balance: $2,400 | Loan balance: $3,250 | | Loan balance: $4,500 |
| Income: $52k | Income: ? | | Income: ? |
| Own House: Yes | Own House: Yes | | Own House: Yes |
| Other delinquent accts: 2 | Other delinquent accts: 2 | | Other delinquent accts: 3 |
| Max billing cycles late: 3 | Max billing cycles late: 4 | | Max billing cycles late: 6 |
| Profitable customer?: ? | Profitable customer?: ? | | **Profitable customer?: No** |
| ... | ... | | ... |

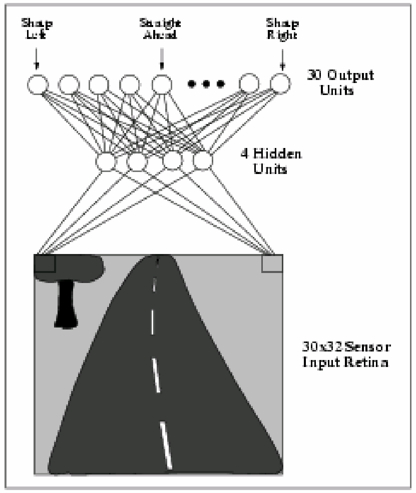
Rules learned from synthesized data:

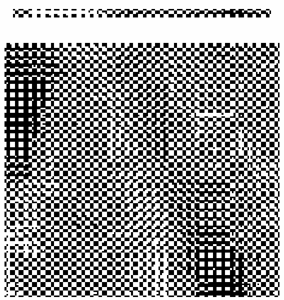
```
If   Other-Delinquent-Accounts > 2, and
     Number-Delinquent-Billing-Cycles > 1
Then Profitable-Customer? = No
     [Deny Credit Card application]

If   Other-Delinquent-Accounts = 0, and
     (Income > $30k)  OR  (Years-of-Credit > 3)
Then Profitable-Customer? = Yes
     [Accept Credit Card application]
```

# Example: Software Applications

- Problems too difficult to program by hand

ALVINN [Pomerleau] drives 70 mph on highways



# Example: Software Applications

- Speech Interface

# Example: User Customization


http://www.wisewire.com

# What is the Learning Problem ?

- Learning=Improving with experience at some task
  - Improve over task T,
  - With respect to performance measure P,
  - Based on experience E

- E.g., Learn to play checkers
  - T: Play checkers
  - P: % of games won against opponents / in world tournament
  - E: opportunity to play against self

# Learning to Play Checkers

- T: Play checkers
- P: Percent of games won in world tournament

- What experience ?
- What exactly should be learned ?
- How shall it be represented ?
- What specific algorithm to learn it ?

# Type of Training Experience

- Direct or indirect ?
  - Direct: board states and the correct move for each
  - Indirect: move sequences and final outcomes

- Teacher or not ?

- Problem: Is training experience representative of performance goal ?

Machine learning rests on the critical assumption that the distribution of training examples is identical to the distribution of test examples

# Choose the Target Function

- *ChooseMove* : *Board* $\rightarrow$ *Move*
- *V* (evaluation function) : *Board* $\rightarrow$ *R*
  - For example,
    - if *b* is a final board state that is won, then *V*(*b*)=100
    - if *b* is a final board state that is lost, then *V*(*b*)=-100
    - if *b* is a final board state that is drawn, then *V*(*b*)=0
    - if b is a not a final state in the game then *V*(*b*)= *V*(*b'*) where *b'* is the best final board state that can be achieved starting from *b* and playing optimally until the end of the game

This gives correct values but is not operational (usable) !

Find an operational description of the ideal target function

- function approximation

# Choose Representation for Target Function

- A table with a distinct entry specifying the value for each distinct board state
- A collection of rules matching against features of the board
- A polynomial of predefined board features
- Artificial neural network

The more expressive the representation,
the more training data will require

# A Representation for Learned Function

- Target function: a linear combination of board features

$$\hat{V}(b) = w_0 + w_1 \cdot bp(b) + w_2 \cdot rp(b) + w_3 \cdot bk(b) + w_4 \cdot rk(b) + w_5 \cdot bt(b) + w_6 \cdot rt(b)$$

  - *bp*(*b*): number of black pieces on board *b*
  - *rp*(*b*): number of red pieces on board *b*
  - *bk*(*b*): number of black kings on board *b*
  - *rk*(*b*): number of red kings on board *b*
  - *bt*(*b*): number of red pieces threatened by black (i.e., which can be taken on blacks next turn)
  - *rt*(*b*): number of black pieces threatened by red

Reduce the problem of learning checkers strategy to the problem of learning values for the weights in the target function representation

# Obtain Training Examples

- To learn the target function $\hat{V}$ we require a set of training examples $\{\langle b, V_{train}(b)\rangle\}$, each describing
  - A specific board state *b* and the training value $V_{train}(b)$
  - Indirect learning is employed

- One rule for estimating training values

$$V_{train}(b) \leftarrow \hat{V}(Successor(b))$$

  - Assumption: values of board states closer to game's end are more accurate

# Choose Weight Tuning Rule

- One common approach is to minimize the squared error between the training values and the values predicted by the hypothesis

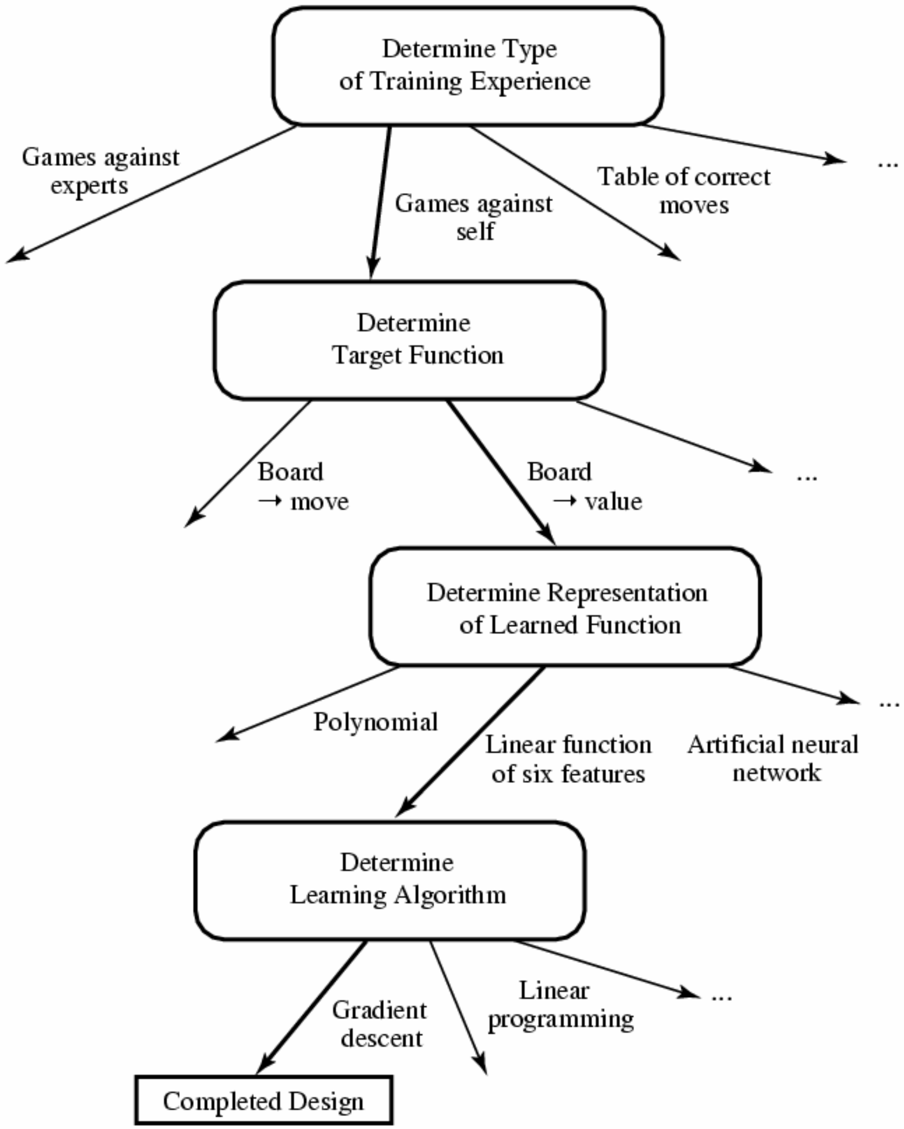
$$E \equiv \sum_{\langle b, V_{train}(b) \rangle \in training \quad examples} \left( V_{train}(b) - \hat{V}(b) \right)^2$$

- Require an algorithm that can
  - Incrementally refine the weights as new training examples become available
  - Robust to errors occurred in the estimated training values

- E.g., gradient-descent search (LMS weight update rule)
  - Repeatedly select a training example *b* at random
    - Use the current weights to calculate $\hat{V}(b)$
    - For each board feature $f_i$ , update the weight $w_i$

$$\tilde{w}_i \leftarrow w_i + \left( V_{train}(b) - \hat{V}(b) \right) f_i$$

# Design Choices




indirect

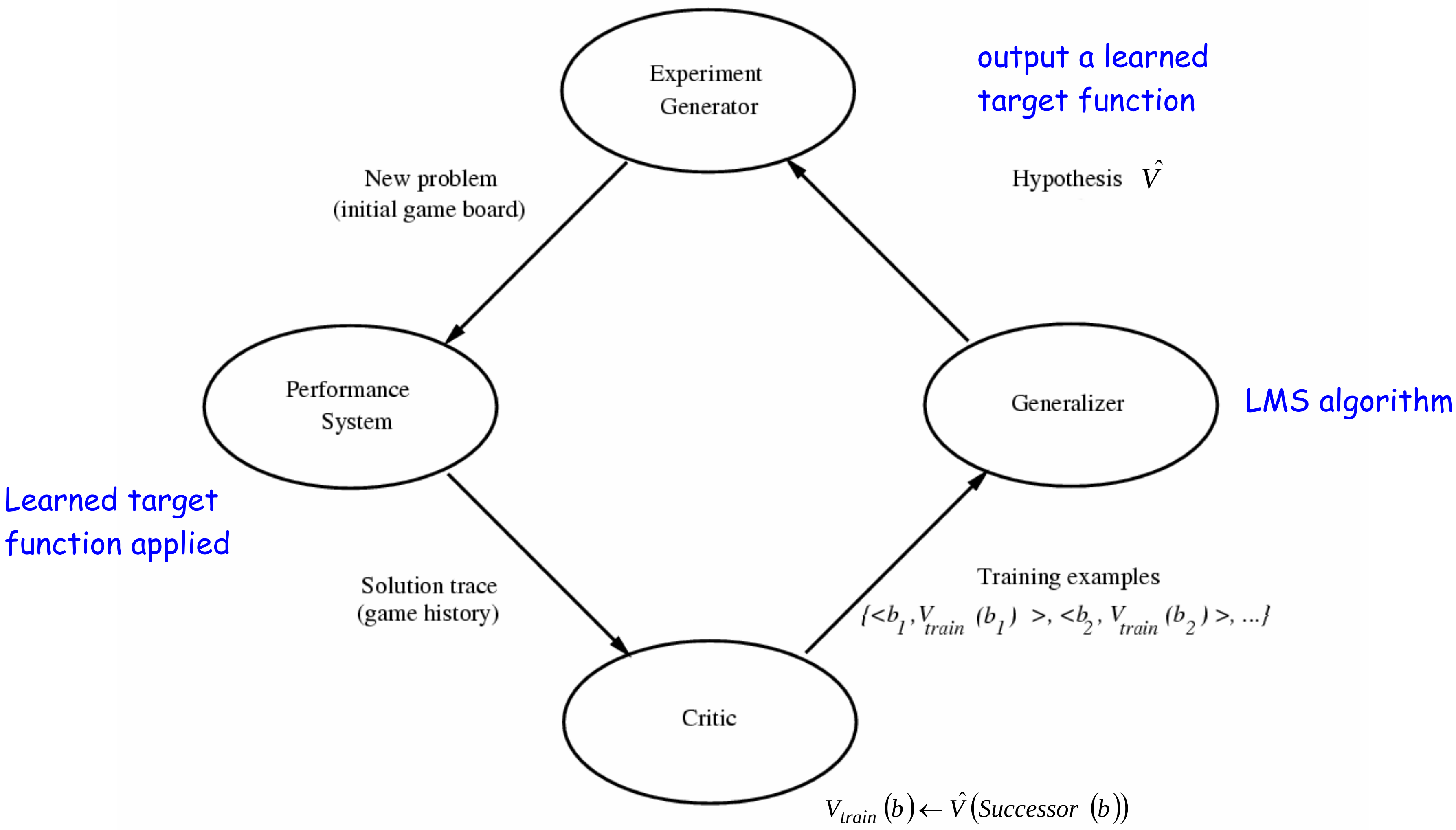
without an external teacher

board state values

linear function

Lest Mean Squares algorithm

# Design of Checkers Learning System



# Some Issues in Machine Learning

- What algorithms can approximate functions well (and when) ?
- How does number of training examples influence accuracy ?
- How can prior knowledge of learner help ?
- How does complexity of hypothesis representation impact it ?
- How does noisy data influence accuracy ?
- What are the theoretical limits of learnability ?
- How can systems alter their own representations ?

# What is Data Mining ?

- Also called *Knowledge Discovery in Databases* (KDD), *Information Extraction* (IE), *Knowledge Extraction* (KE) ..
- Emerged during the late 1980s, has made great strides during the 1990s, and continues to flourish into the new millennium

# What is Data Mining ?

- Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
  - Extract/Mine interesting information or knowledge (rules, regularities, patterns, constraints) from huge amounts of data stored in databases, data warehouse, and other information repositories
  - “knowledge mining” from data

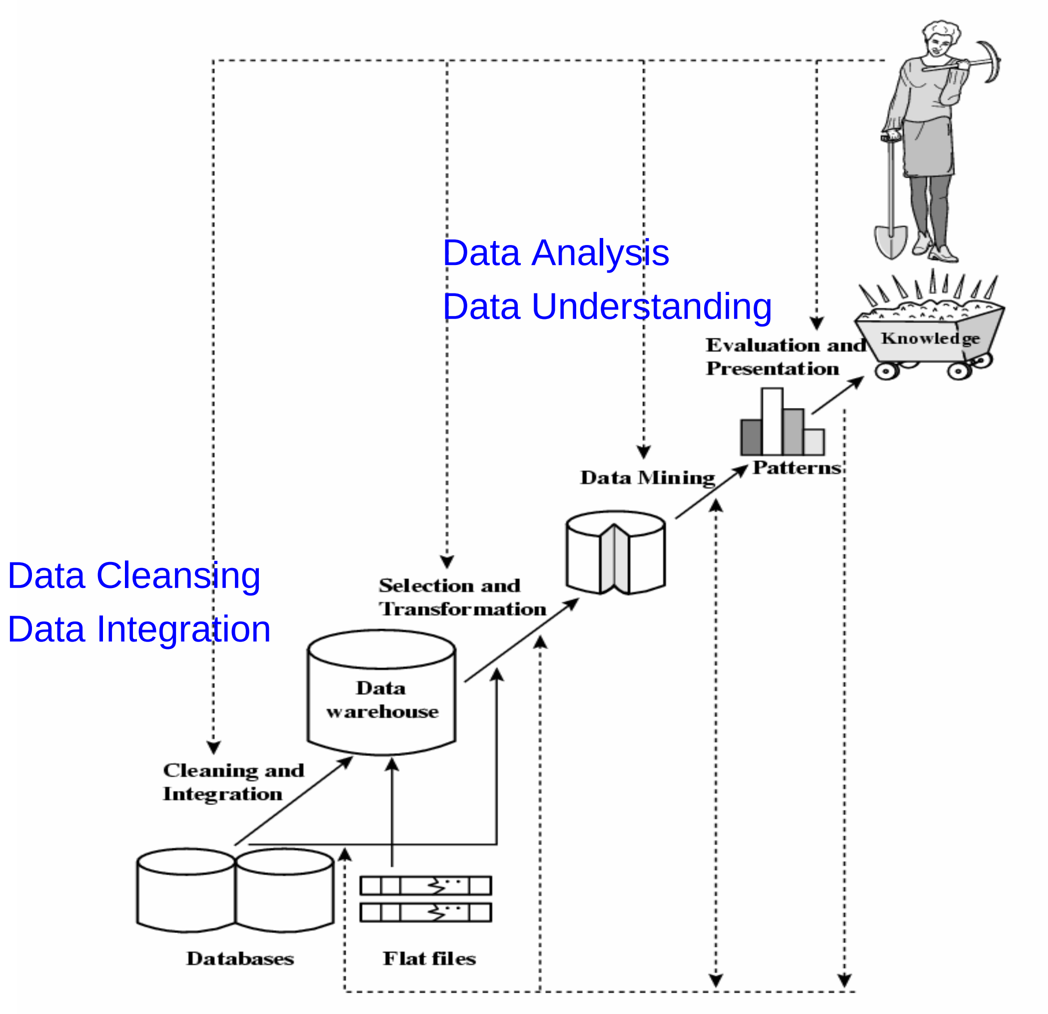
# What is Data Mining ?

Data Mining Data Mining

Data → Information → Knowledge


# What is Data Mining ?

- Data mining is an essential step in knowledge discovery



# Categories of Data Mining

- Predictive Data Mining
  - Produce the model of the system described by the given data set
  - I.e., perform inference on the current data to make predictions
    - Classification
    - Regression

- Descriptive Data Mining
  - Produce new, nontrivial information (uncover patterns and relationships) based on the available data set
  - I.e., characterize the general properties of the data
    - Clustering
    - Summarization, or Concept/Class Description
    - Dependency/Association Modeling $age(X, \text{"20...29"}) \wedge Income(X, \text{"20K...29K"}) \Rightarrow Buy(X, \text{"CD Player"})$
    - Change and Deviation Detection evolution, outlier detection

# Multi-Dimensional View of Data Mining

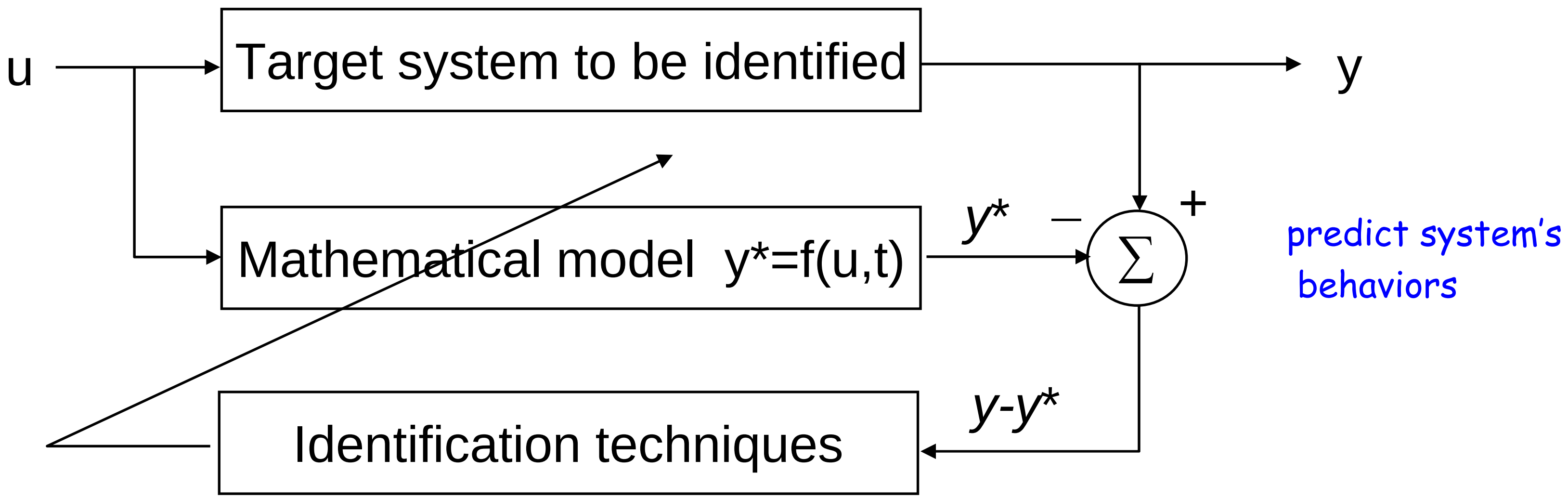
- Databases to be mined
  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.
- Knowledge to be mined
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Granularity: mining at multiple levels of abstraction
- Techniques utilized
  - Machine learning, statistics, visualization, neural network, database-oriented, data warehouse (OLAP), etc.
- Applications adapted
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc

# Roots of Data Mining

- Statistics, Mathematics
  - Models
- Machine Learning
  - Algorithms
- Control theory
  - System identification

# Roots of Data Mining

- System Identification (an iterative process)
  – Structure Identification
  – Parameter Identification



# Phases of Data Mining

1. State the Problem and Formulate the Hypothesis
   - The problem statement should be established based on domain-specific knowledge and experience
   - But application studies tend to focus on the data-mining technique at the expanse of a clear problem statement
   - Cooperation between data-mining expertise and application expertise

# Phases of Data Mining

2. Collect the Data

- Two possible approaches
  - Designed experiment
    - Data generation process is under control of an expert
  - Observational approach (random data generation)
    - The expert can not influence the data generation process

- A prior knowledge can be very useful for modeling and final interpretation of results

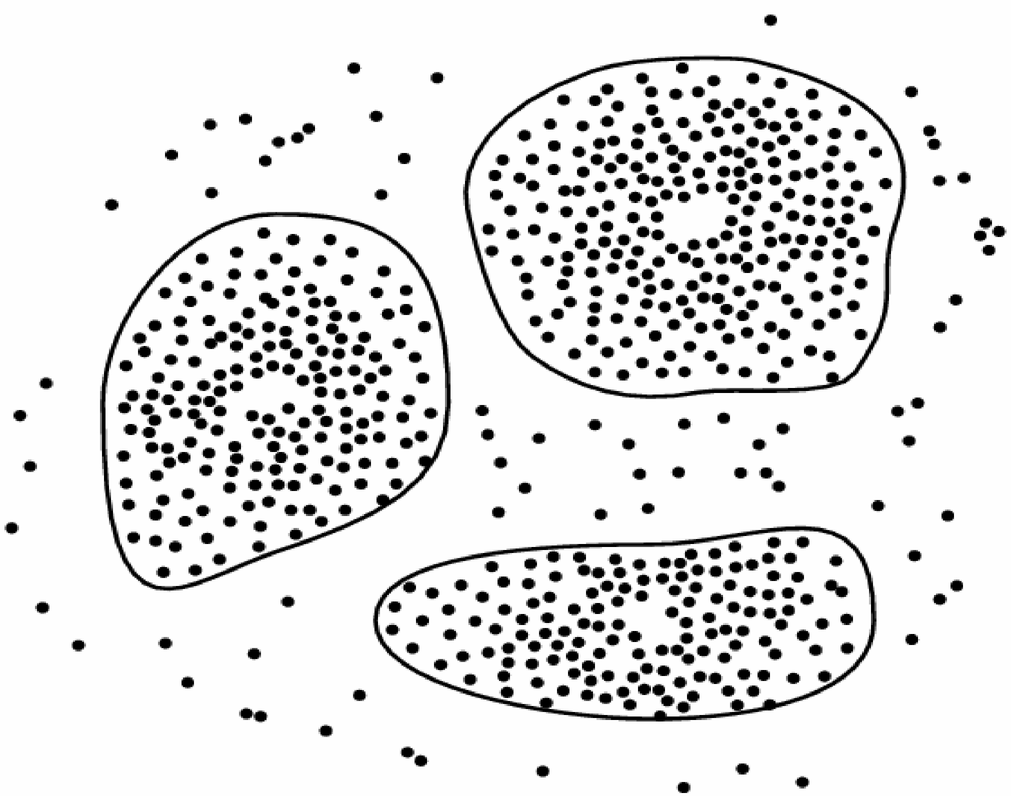
- Data respective for estimating a model and testing should come from the same, unknown, sampling distribution

# Phases of Data Mining

3. Preprocessing the Data
   - Two tasks involved
     - Outlier detection (and removal)
       - Outliers are unusual data values that are not consistent with most observations which can seriously affect modeling accuracy
       - Two strategies for dealing with outliers
         - » Removal of outliers
         - » Robust modeling methods
     - Scaling, encoding, and selecting features (dimensionality reduction)
   - The prior knowledge of application domain should be considered in data-preprocessing steps

# Phases of Data Mining

- Clusters and Outliers



# Phases of Data Mining

4. Estimate the Model
   - Select and implement the appreciate data-mining technique
     - The implementation is based on several models

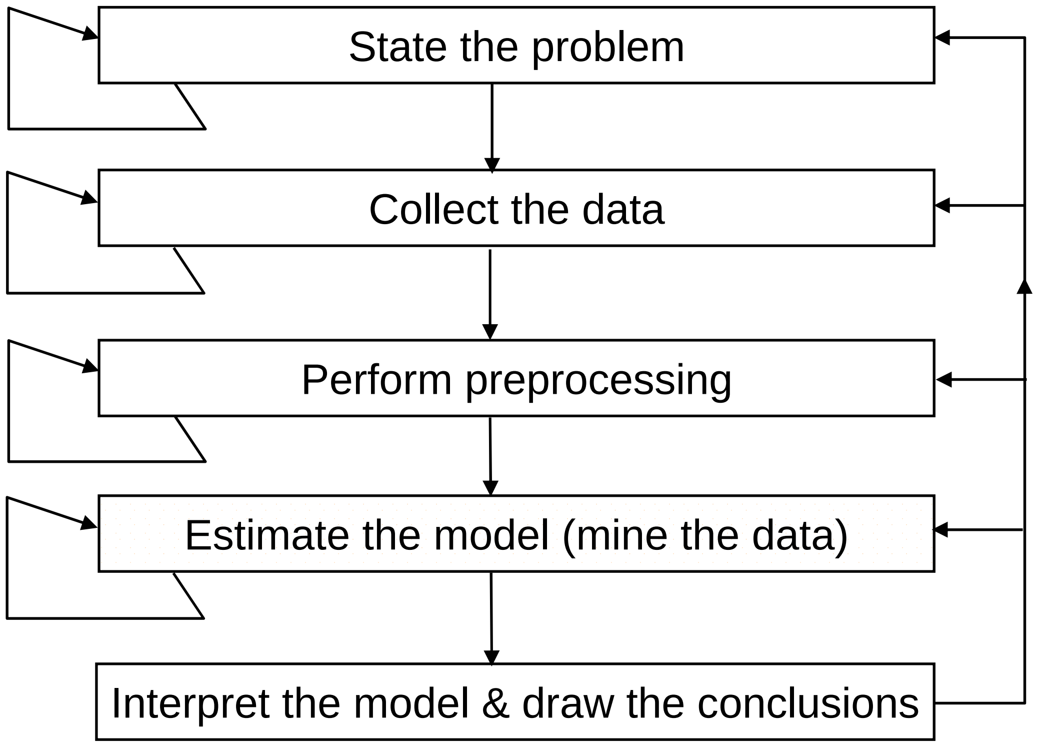
   - Use the technique to learn and discovery information from large volumes of data sets

# Phases of Data Mining

5. Interpret the Model and Draw Conclusions
   - Data-mining models should help in decision making
   - Data-mining models thus should be interpretable
   - Tradeoff between accuracy of model and accuracy of model's interpretation

# Phases of Data Mining

- All phases and the entire data-mining process are highly iterative



# Large Data Sets

- An exponential growth in information sources and information-storage units

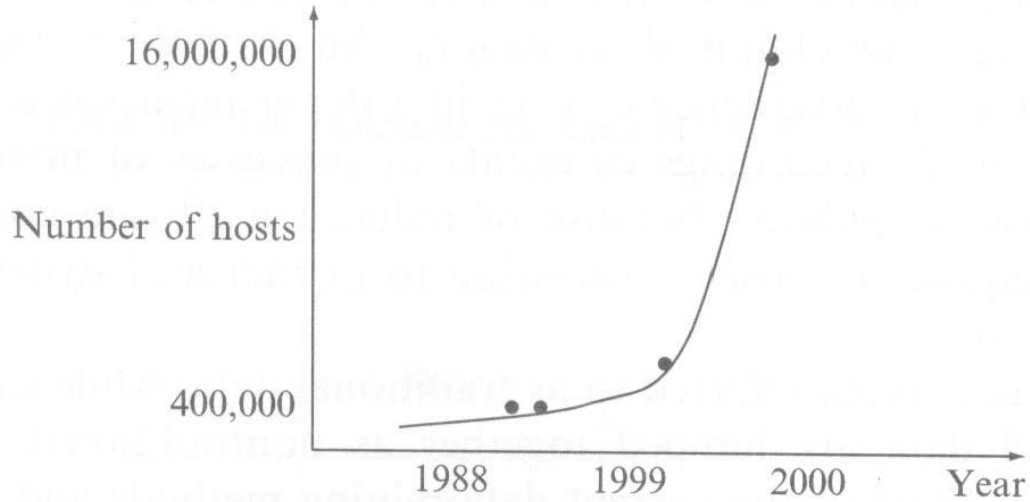


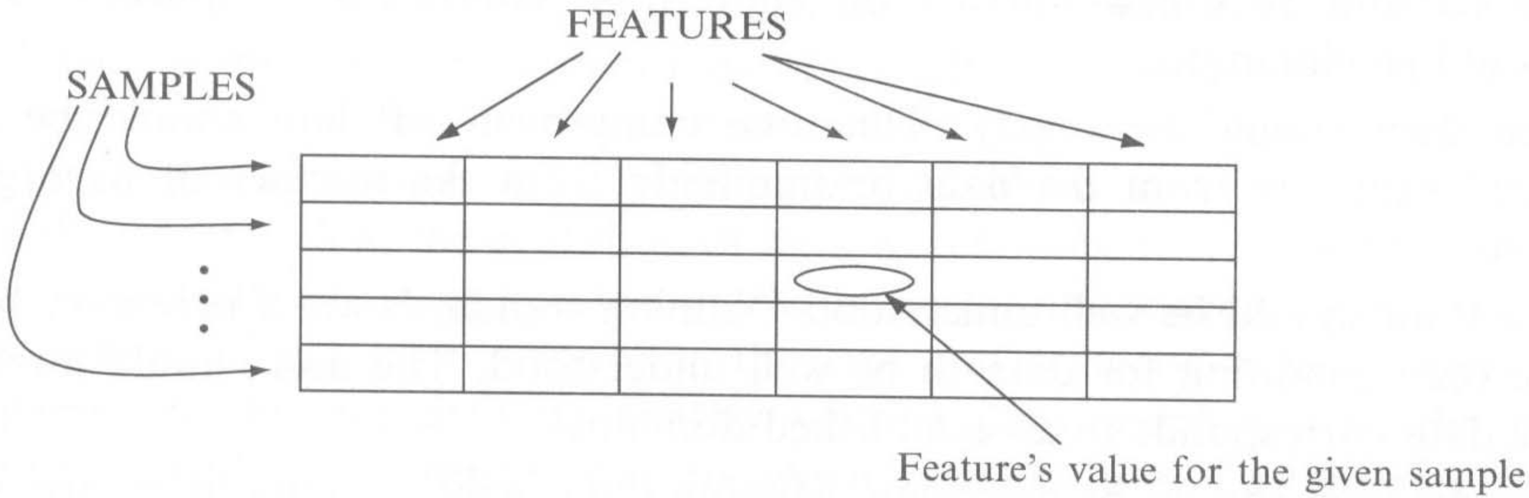

**FIGURE 1.3** Growth of Internet hosts

  – The number of hosts are directly proportional to the amount of data stored on the Internet

# Large Data Sets

- Infer knowledge form huge volumes of raw datasets
  - Big data can lead to much stronger conclusions
  - A rapidly widening gap between data-collection and data-organization capabilities and the ability to analyze the data
  - Manual analysis and semiautomatic computer-based analysis can not deal with the large volumes of data sets
- Data as the sources for data mining can be classified into structured, semi-structured and unstructured data
  - Traditional data: structured data
  - Nontraditional data (multimedia):: semi-structured and unstructured data

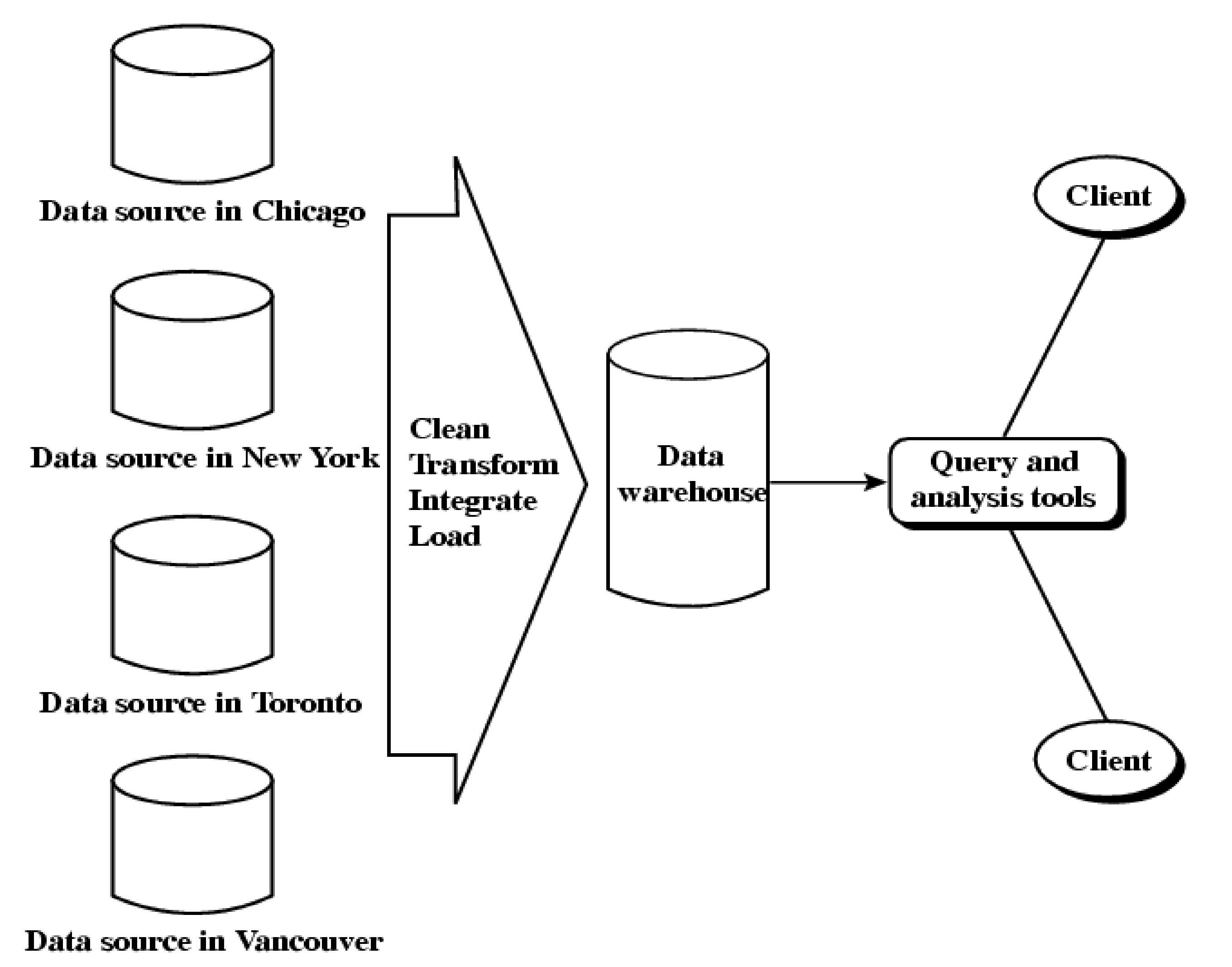
# Structured Data



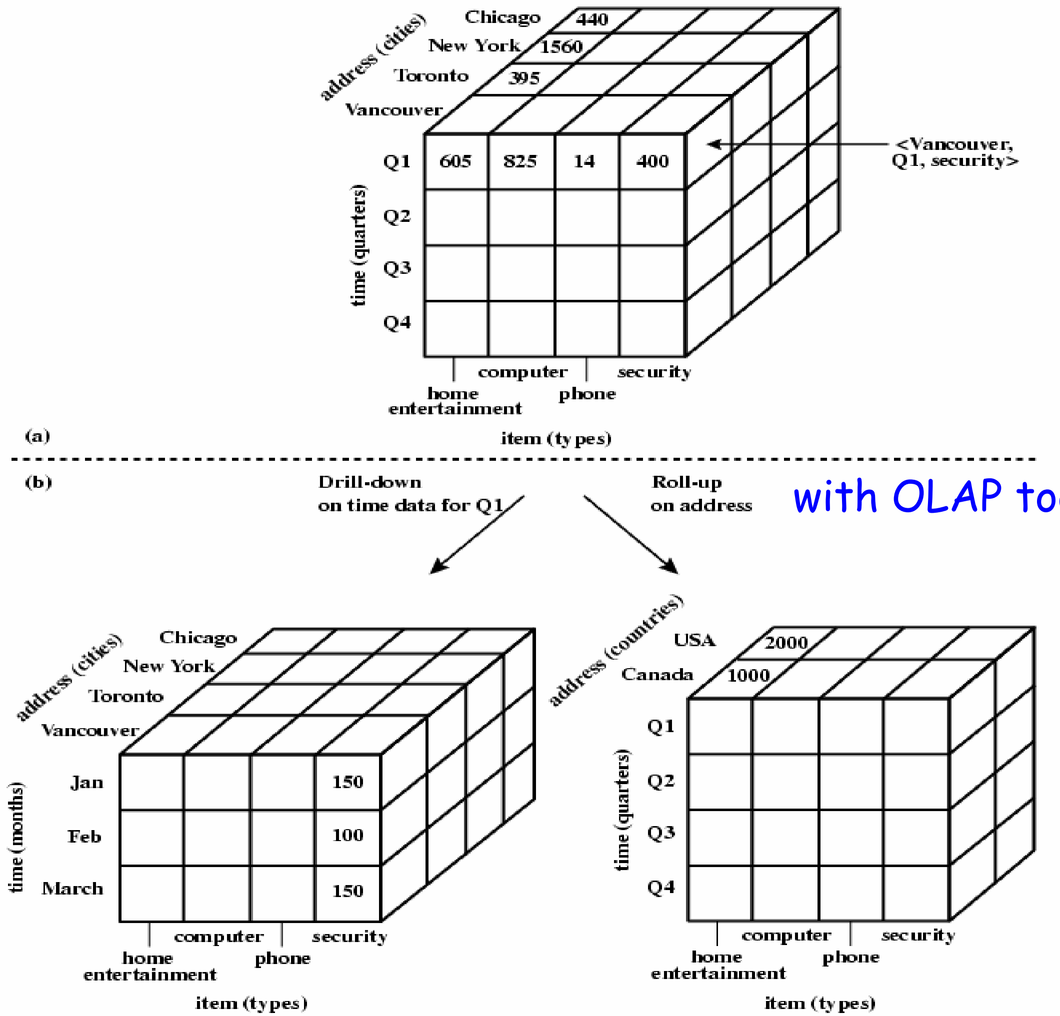
# Data Warehouse

- Definition
  - A collection of integrated, subject-oriented databases designed to support the decision-support functions (DSF), where each unit of data is relevant to some moment in time
    - Modeled as a multidimensional database structure
  - Or, a repository of multiple heterogeneous data sources, organized under a unified schema usually at a single site in order to facilitate management decision making
- That is, the sole of a data warehouse is to provide information for end users for decision support
- Cf. data mart
  - A department subset of a data warehouse

# Data Warehouse



# Data Warehouse




with OLAP tools

# Data Warehouse Applications

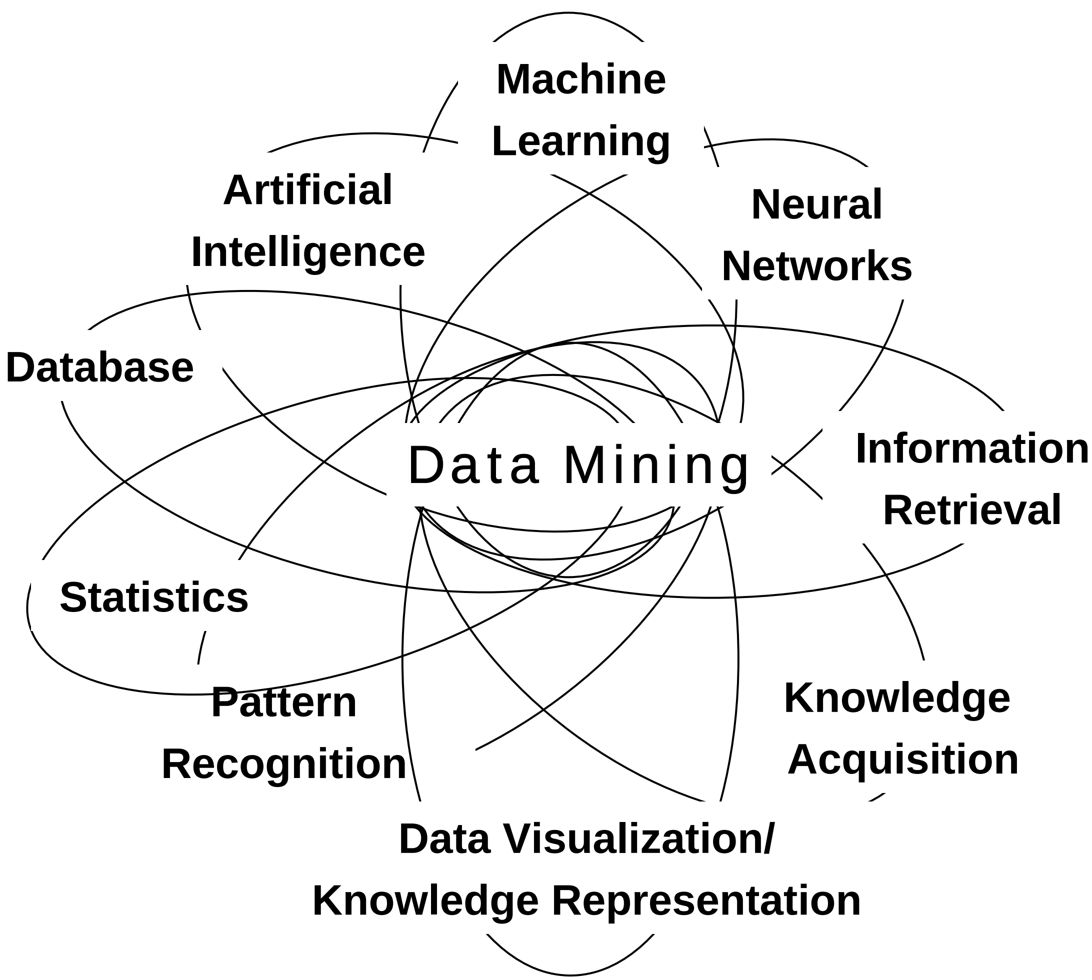
- Data mining
  - Represent one of the major applications for data warehouse
  - Provide end-user with the capability to extract hidden, nontrivial (not obvious) information
    - Act as exploratory queries
- Structured query languages (SQL)
  - A standard database language
  - Used when we know exactly what we are looking for and we can describe it formally
- Online Analytical Processing (OLAP)
  - Do not learn from data, nor create new knowledge
  - Let users analyze data by providing multiple views of the data

# Data Warehouse

- Classification of data stored in a data warehouse
  - Old detail data
  - Current (New) detail data
  - Lightly summarized data
  - Highly summarized data
  - Metadata (the data directory or guide)

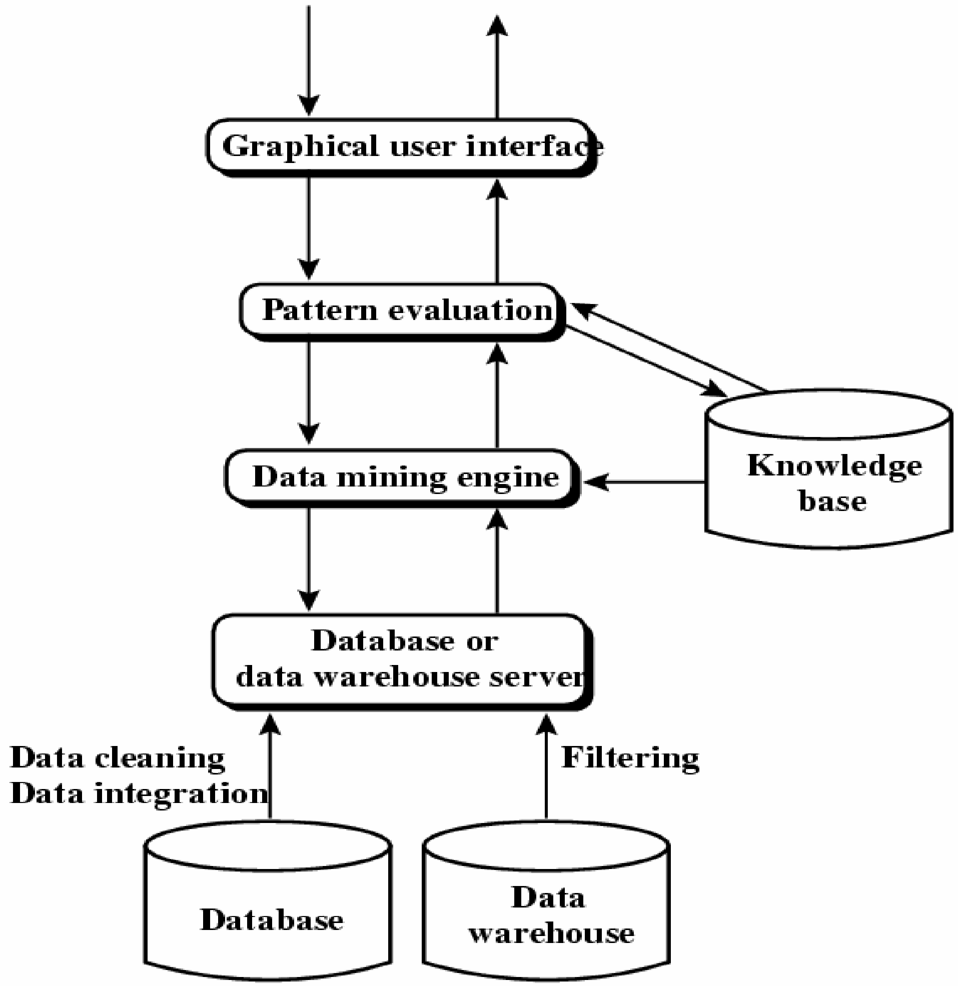
- Fundamental types of data transformation
  - Simple transformations (encoding/decoding)
  - Cleansing and scrubbing
  - Integration
  - Aggregation and summarization

# Confluence of Multiple Disciplines



# A Typical Data Mining System

- Architecture



# Course Topics

- Data Preparation and Data Reduction
- Concept Learning
- Cluster Analysis
- Decision Trees and Decision Rules
- Statistical Learning Theory
- Bayesian Learning and Related Statistical Learning Methods
- Association Rules
- Reinforcement Learning
- Hidden Markov Models
- Artificial Neural Networks
- Genetic Algorithms
- Support Vector Machines

# Topic List and Schedule

| | | | |
|---|---|---|---|
| 2/19 | | Course Overview & Introduction | |
| 2/26 | | Concept Learning (ML Ch. 2) | |
| 3/4 | | Data Preparation and Data Reduction (DM Ch. 2~3) | |
| 3/11 | | Decision Trees and Decision Rules (ML Ch. 3, DM Ch. 7) | |
| 3/18 | | **Break ?** (ICDAT 2004) | |
| 3/25 | | Cluster Analysis (DM Ch. 6) | |
| 4/1 | | Bayesian Learning (ML Ch. 6) | |
| 4/8 | | Statistical Learning Theory and Statistical Methods (DM Ch. 4~5) | |
| 4/15 | | **Midterm** | |
| 4/22 | | Artificial Neural Networks (ML Ch. 4, DM Ch. 9) | |
| 4/29 | | Association Rules (DM Ch. 8) | |
| 5/6 | | Genetic Algorithms (ML Ch. 9, DM Ch. 10) | |
| 5/13 | | Evaluating Hypotheses (ML Ch. 5) | |
| 5/20 | | **Break** (ICASSP 2004) | |
| 5/27 | | Learn Sets of Rules (ML Ch. 10) | |
| 6/3 | | Support Vector Machine, Hidden Markov Models | |
| 6/10 | | Reinforcement Learning | |
| 6/17 | | **Final Exam** | |

# Journals & Conferences

- Journals
  - *Machine Learning*
  - *IEEE Transactions on Pattern Analysis and Machine Intelligence*
  - *Neural Networks*
  - *…..*

- Conferences
  - *International Conference on Machine Learning*
  - *International Conference on Knowledge Discovery and Data Mining*
  - *……*